How Goals Shape Category Acquisition: The Role of Contrasting Categories

Tyler Davis (thdavis@mail.utexas.edu)
Bradley C. Love (brad_love@mail.utexas.edu)
Department of Psychology, The University of Texas at Austin
Austin, TX 78712 USA

Abstract

Categories associated with goals are often organized around ideals rather than central tendencies. In such real-world categories, items that are extreme are often perceived as being the most typical. Here, we demonstrate similar effects with artificial categories learned in the laboratory. Our experiment and simulations suggest that low-level learning mechanisms that seek to minimize prediction error may be responsible for aspects of category idealization. To minimize error, category centroids are adjusted to both increase similarity to their members, as well as to minimize similarity to members of contrasting categories. This learning dynamic distorts category representations away from contrasting categories, leading to idealization.

Keywords: Category learning, category use, goals.

Introduction

Increasingly, research suggests that a learner's goals within an environment place substantial constraints on the content of their category representations (for review see Markman & Ross, 2003). For example, categories learned by classification lead to important differences between those learned from inference, even though the two tasks are formally equivalent. These effects that goals have on category representations provide a critical challenge for formal approaches to category learning, as many models do not contain ways in which a learner's goals can exert any influence over their representations (cf. Love, 2005). At the same time, because of this challenge, goals offer a substantial opportunity for cross-fertilization between fields such as those studying expertise/cultural psychology and those studying formal modeling/laboratory experimentation. In the following, we discuss ways in which goal related influences on representations can be studied using formal models, and design a novel experiment using predictions from the model and previous results from the cultural/expertise literatures.

One finding in the cultural, expertise, and goal-derived category literatures is that categories associated with goals have a *graded structure* determined by ideals. Graded structure refers to the notion that members of a category differ in terms of how typical they are for the category to which they belong. Whereas many categories have a graded structure determined by the category's statistical central tendency (Rosch, 1975), goal-derived categories have a graded structure centered around the category ideal (Barsalou, 1985). For example, the goal derived category *foods to eat on a diet* will have a graded structure near 0 calories (e.g., celery). Ideals have also been shown to influence the graded structure of taxonomic categories when groups or other cultures that have goals associated with the category are tested (e.g., Atran, 1999). For example, tree experts have been shown to view trees as more typical to the extent that they minimize weediness (Lynch, Coley, & Medin, 2000).

The majority of explanations for how this *idealization* occurs suggest that these effects depend upon real-world and cultural influences that are not present in artificial categories commonly used in the laboratory. Thus, these findings are often depicted as being at odds with approaches to categorization that rely on laboratory techniques such as mathematical modeling. It is our perspective, however, that these methods are continuous with one another, and that they can be mutually informative. Further, we'll argue that lower level learning mechanisms described by e.g., error-driven clustering models, provide at least a partial explanation for how goals affect graded structure, whilst making predictions for how these effects may be demonstrated in the laboratory.

A Formal Model

Clustering models can be used to place a different perspective on how idealization occurs in categories defined by goals. To illustrate, in the diet food example above, a clustering model would have separate clusters representing diet foods and non-diet foods that would initially be centered on the mean number of calories in their respective categories. However, since the distributions of calories in these two categories are continuous, and even somewhat overlapping, sometimes food from one category will highly activate the cluster of the other category leading to error. Error-driven clustering models will compensate for this error by moving the cluster means on the calorie dimension further away from each other. After some trials, the clusters will tend to move further apart than the actual category means, producing a marked shift in the graded structure of the categories. Whereas at the outset of learning category members that are near the statistical central tendency of the category will produce the strongest activation, at the end of learning, items that are more extreme (i.e., ideal) on the goal dimension will.

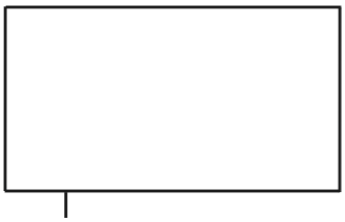

Figure 1- An example stimulus. The dimensions of variation are the height and the position of the line segment.

To illustrate how this happens in an error-driven clustering model, we will use a simple model that is able to capture this overall pattern, but is not a full model of category learning in and of itself. This will allow us to describe a set of principles that underlie the entire class of error-driven clustering models, without having to deal with the added complexity of attentional learning and cluster recruitment that necessarily accompany these models in their more complete forms.

Formally, the model represents each category as a cluster that gives the category's mean and the standard deviation along each dimension for which it is defined. Each time a stimulus is encountered, activations are computed for each cluster, and the strength of these activations determines how the stimulus will be classified. Activation, a, for a given cluster i is given as a Gaussian function of the presented stimulus value j:

$$a_i = \frac{1}{s_i\sqrt{2\pi}} e^{-\frac{dij^2}{2si^2}} \qquad (1)$$

Where si is the cluster's standard deviation (this value is constant in the present application), and d_{ij} is the distance between presented stimulus j and the position of cluster i given by:

$$d_{ij} = \sqrt{\sum_m (x_{jm} - x_{im})^2} \qquad (2)$$

Where x_{jm} is the value of stimulus j on dimension m, and x_{im} is the mean of cluster i on dimension m.

Cluster means are learned by gradient descent on an error, which is the mechanism that allows the graded structure of the categories to approach the category ideals. Whenever the position of a cluster along a dimension causes the cluster to become too highly activated in response to a non-category member, the following equation updates the mean along this dimension in proportion to the magnitude of the error.

$$E = \frac{1}{2}(t_i - a_i)^2 \qquad (3)$$

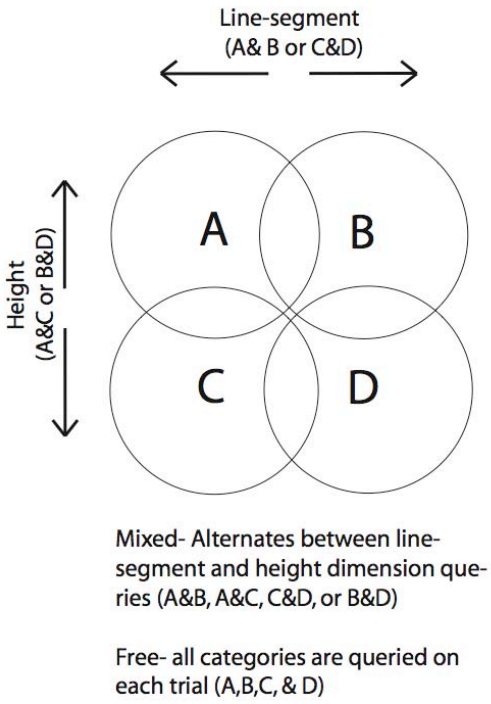

Figure 2 – Category structure and conditions. The letters in parentheses give the possible choices on a given type of trial within a condition.

$$\Delta x_{im} = -\lambda \frac{\partial E}{\partial M} = \lambda(t_i - a_i)\frac{x_{jm} - x_{im}}{s_i^3\sqrt{2\pi}} e^{-\frac{dij^2}{2si^2}} \qquad (4)$$

Where λ is a learning rate for the cluster mean, ti is the feedback to cluster i,

$$t_i = \max(alpha, a_i), \text{ if the item is in the category corresponding to cluster i}$$
$$t_i = \min(0, a_i), \text{ if the item is not in the category corresponding to cluster i.} \qquad (5)$$

In the applications described below, alpha, λ, and si are all free parameters.

Experiment and Predictions

Translating the model's predictions into ones that can be tested using artificial categories in the laboratory, this suggests that idealization occurs in cases that require discriminating between stimuli along particular dimensions (see also Goldstone, 1996). In this experiment, we create a category learning task that mimics this property of the goal-derived categories described above. The stimuli for this experiment vary along two continuous dimensions (see Figure 1), and are partitioned into four categories that are separated in different respects along these dimensions (see Figure 2).

Different discrimination goals are made salient in this experiment by varying, between conditions, the types of contrasts subjects use to learn the categories, while keeping the categories themselves constant. In the two unidimensional conditions (labeled Line-segment and Height in Figure 2), subjects learn the categories by contrasting those that vary on only a single dimension. On any given trial in these conditions, subjects will have the option of choosing between categories that are discriminable on only one of the two dimensions that the stimuli are defined on (e.g. A & B or C & D in the line-segment condition or A & C or B & D in the height condition). In two other conditions, both dimensions will be made relevant either by alternating between trials that allow subjects to choose categories separated on either the line-segment or height position (label Mixed in Figure 2), or allowing subjects to choose freely between all categories on a given trial (labeled Free in Figure 2).

The model's predictions for these conditions are straightforward and analogous to the discussion of how the model would apply to diet/non-diet food contrasts. In any condition for which a dimension is relevant to learning the categories, the clusters will move away from each other on this dimension as a result of the error-driven learning mechanisms. However, this will not be the case for the irrelevant dimension in these conditions. Since irrelevant dimensions do not help to discriminate between the categories, the model will predict no shift in graded structure along these dimensions.

To assess whether subjects' graded structure for these categories has changed as a result of this learning, we will have them reconstruct average stimuli from each category, and provide typicality ratings for stimuli observed during the category learning task. If the graded structure of these categories has been affected by goals, as predicted by the model, these reconstructions should be shifted away from the true mean of the categories and away from the categories that they have been contrasted with. Similarly, the typicality ratings should show that items that are further away from opposing categories on the dimensions that were contrasted during learning are more typical than items that are closer to the statistical average of the category along these dimensions.

It is important to note that these predictions derived from the model are not predicted by clustering models without error-driven learning (e.g., Anderson, 1991) or classical Roschian accounts of graded structure (Rosch, 1975). These accounts predict that the reconstructed category averages and typicality ratings will not depend on how the categories were contrasted during learning. Instead, all conditions would be expected to have the same graded structure determined by the categories' statistical central tendency. Further, Bayesian accounts of memory for stimulus magnitude predict that the reconstructed averages would actually distort toward the center of the overall category distribution (Huttenlocher, Hedges, & Vevea, 2000; Sailor & Miram, 2005). This is opposite from that predicted by error-driven clustering models.

Method

Subjects 188 students from the University of Texas at Austin participated for course credit.

Stimuli Stimuli were blue rectangles that varied in terms of their height and the position of a vertical line segment along their lower base. The rectangles had a fixed width of 60mm, and their height and line segment position were sampled on each trial from one of four category distributions (A, B, C, D). These distributions were approximately normal with standard deviations of 2.4mm and centered on (15mm, 21mm), (21mm, 21mm), (15mm, 15mm), (21mm, 15mm) from the left side of the rectangle and base respectively. To keep the absolute range a category was allowed to vary over constant and allow for some overlap between categories, all stimuli were required to be within 2 standard deviations of their respective mean.

Design Subjects completed a category learning phase, followed by a reconstruction phase in which they produced the average of each category.

The category learning phase required subjects to learn four categories: A, B, C, and D (see Figure 2). Conditions differed in how they were required to learn these categories in that subjects were given a restricted set of possibilities on each trial for which category a stimulus belonged. In the unidimensional conditions, subjects only had to decide between categories differing on the line segment dimension (A&B or C&D) or on the height dimension (A&C or B&D), but not both. The mixed condition varied these two types of comparisons, and subjects in the free condition were allowed to decide between all possible categories on every trial. Category labels were randomized for each subject.

In order to complete the category learning phase, subjects were required to either achieve a criterion of 80% correct in a single block of 40 stimulus presentations (10 from each category), or complete 5 blocks total. On each trial a stimulus was sampled at random. The mean height and line segment position of stimuli in a given category in a given block were required to equal the mean of the distributions that they were drawn from.

In the reconstruction phase, subjects were asked to recreate the average member of each category. To do this, they used the arrow keys to adjust the height and line segment position of a stimulus randomly drawn from the respective category distribution. This phase lasted for three blocks in which each category was queried once. Trial order was randomized with the constraint that the same category could not appear twice in a row.

In the typicality phase, subjects were asked to rate how typical each stimulus was on a continuous scale that was

presented in red below the stimulus. Stimuli for the typicality phase were the same stimuli from the first two blocks of the category learning phase.

Procedure Directions were displayed on the screen prior to each phase. Subjects wore headphones to deliver auditory feedback and dampen background noise.

On each category learning trial a stimulus was presented in the center of the screen along with a prompt telling the subject which categories they could choose from. After keying in their response, they were given corrective feedback for 2000 ms followed by a blank screen for 500 ms.

On each trial of the reconstruction phase, a stimulus was presented in the center of the screen along with a prompt telling the subject which category average to adjust the stimulus to resemble. The subject used the arrow keys to continuously adjust the stimulus, and each trial was self-paced. No feedback was given during reconstruction.

On each trial of the typicality phase, subjects were prompted to rate the typicality of a given stimulus for the category to which it belonged, by moving a red dot on a line with the arrow keys between ends marked "Very Typical" and "Not Typical." Each trial was self pace, and no corrective feedback was given.

Results and Discussion

Category Learning In the present experiment, the category learning phase was meant to serve as the key manipulation in predicting how subjects would respond in later portions of the experiment (i.e., the reconstruction and typicality phases). As such, the category learning data was not particularly important for any of the hypotheses we discussed in the introduction in and of itself. Instead, we use the category learning data only as a measure of which participants successfully learned the task, and should therefore be included in the analysis of the reconstruction and typicality data. Subjects were excluded from the rest of the analyses if they failed to perform significantly better than chance in their final classification block. However, because the number of available choices on each trial differed between conditions, chance also differed. In conditions 1, 2, and 3 subjects were excluded if they answered less than 62.5% correct, and in condition 4 where four categories could be chosen from on a given trial, subjects were excluded if they answered less than 37.5% correct. These values were based on binomial distributions with p=.5 (conditions 1, 2 & 3) or p=.25 (condition 4), N=40, and 95% confidence levels. This resulted in the removal of 12 subjects in the line segment unidimensional condition, 5 subjects in the height unidimensional condition, 11 subjects in the mixed condition, 9 subjects in the free condition. Removing these subjects did not affect the pattern of results.

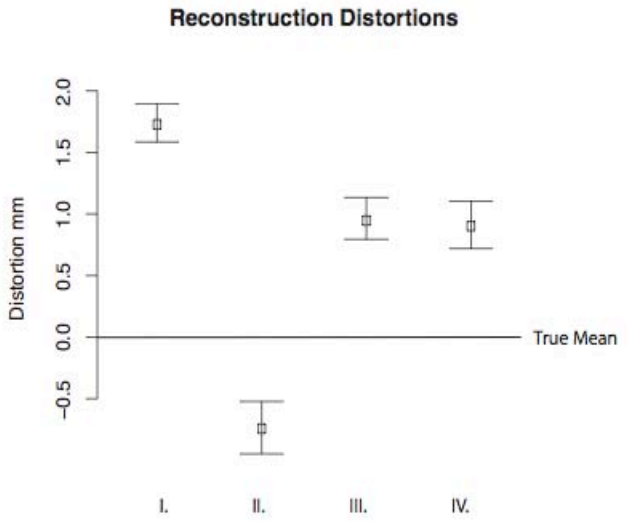

Figure 3- Reconstruction results by condition. I. Unidimensional- relevant dimension II. Unidimensional-irrelevant dimension III. Free IV. Mixed. Positive scores are away from opposing categories and negative scores are toward opposing categories.

Reconstruction Reconstructions were scored as distortions from the respective category mean, with negative scores depicting distortions toward the center of the stimulus space (i.e., toward the opposing categories), and positive scores depicting distortions away from the center of the stimulus space (away from the opposing categories). Scores were calculated for each dimension for each individual subject. Because differences between physical dimensions were not of interest, we pooled the reconstruction results into four data points: rule-relevant dimension reconstructions in both unidimensional conditions (line segment & height), irrelevant dimension reconstructions, both dimensions in the mixed condition, and both dimensions in the free condition. See Figure 3 for reconstruction results.

In the unidimensional conditions on the relevant dimension, the mean reconstruction was 1.738 mm. This represents a significant distortion away from the true mean of the category (0), $t(76)=11.28, p<.01$, and away from the opposing categories. In contrast, on the irrelevant dimension the mean reconstruction was -.735 mm, indicating that subjects in the unidimensional conditions tended to distort in the opposite direction (toward the opposing categories), $t(76)=3.46, p<.01$. In the other conditions, free and mixed, subjects distorted away from the opposing categories slightly less than in the unidimensional conditions with a mean reconstruction of .963 mm in the free condition $t(35)=5.68$, $p<.01$, and .912 mm in the mixed condition $t(36)=4.76$, $p<..01$. These conditions both distorted away from the opposing categories significantly less than observed in the unidimensional conditions ($p<.01$).

These data show that if a dimension is relevant in learning a task, subjects' reconstructed averages along this dimensions shift away from opposing categories, and are predicted from the current account of shifts in graded structure arising from learning mechanisms. The negative shifts (toward the center of the category distribution) observed on the irrelevant dimension in the horizontal and vertical conditions are not predicted by the present framework. However, they would be predicted by models of memory for stimulus magnitude that show that subjects' reconstructions are biased toward the overall mean in tasks using single categories or where the difference between categories is learned incidentally (Huttenlocher, Hedges, & Vevea, 2000; Sailor & Miram, 2005).

Typicality For analysis of the typicality data, individual stimuli were scored according to their distance from their respective category mean (see reconstruction results) with negative values depicting stimuli that are closer to the center of the stimulus space on a given dimension, positive values depicting stimuli that are closer to the extremes on a given dimension, and zero representing the actual category mean. Figure 4 shows typicality as a function of the position of a stimulus within its respective category for each condition. The typicality graph for the unidimensional conditions are shown with the line segment dimension as the relevant dimension, however the data used to construct the graph are averaged from both the line segment and height conditions.

In order to quantify the differences that are apparent in the graphs, we ran individual regression models on each subject's typicality data using the values of the stimuli along both dimensions as predictors. Positive regression slopes along a dimension indicate that as stimuli become further away from opposing categories along this dimension, the observed typicality increases, whereas negative slopes indicate a decrease. Thus, significantly positive slopes along a dimension indicate that the graded structure is shifted toward the ideals along this dimension.

Significantly positive slopes were observed in the unidimensional conditions for the relevant dimension (.273) $t(76)=14.04$, $p<.01$, and the irrelevant dimension (.047) $t(76)=2.77$, $p<.01$. While the irrelevant dimension was not predicted to have a significant impact on typicality scores, the slopes on this dimension were significantly smaller $t(76)=8.08$, $p<.01$, showing that this impact was less than with the relevant dimension. In the free condition typicality increased positively as the observed value of the stimuli increased on both dimensions (.186) $t(35)=10.11$, $p<.01$, and the same occurred in the mixed condition (.196) $t(35)=10.02$,$p<.01$.

These data largely concur with those hypothesized above for typicality ratings. If a dimension is relevant for learning the categories, as stimuli become more extreme (i.e., ideal) along this dimension, they become more typical.

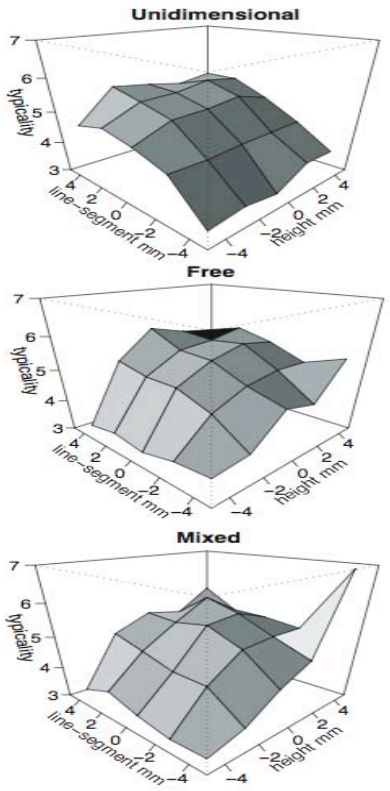

Figure 4- Typicality in each condition as a function of stimulus position.

Model-Based Analysis

In this section, we illustrate how the error-driven clustering model that we introduced above is able to account for the data from the reconstruction phase of the experiment. To accomplish this, the model was trained using the same general procedure as described above for human subjects, except that on each iteration it was trained for the full five blocks and not cut-off upon reaching a criterion. The model's predictions for reconstruction were obtained by using the centered (see reconstruction results) final cluster positions, and hand fitting the model to the average reconstruction on the relevant dimension from the two unidimensional conditions.

As discussed above, idealization effects like those observed in the reconstruction and typicality data are an a priori prediction of error-driven clustering models. As such, there are no parameter settings at which results opposite those obtained could be predicted. Because of this, fitting the model only served to calibrate the predicted cluster distortion to the level observed in the experiment, and also to examine how the model predicts performance across conditions to differ.

The model's predictions were calculated using the average final cluster positions obtained from 10,000 simulations with parameters set at: λ=142.70, alpha=.05, and si= 5. The reconstruction predictions are 1.738 mm

for the unidimensional conditions on the relevant dimension, ~0 mm for on the irrelevant dimension, 2.076 for both dimensions in the free condition, and 1.009 for both dimensions in the mixed condition. The model therefore predicts all of the qualitative effects discussed in the introduction and demonstrated in the reconstruction data.

The model is able to capture the direction of the distortions along relevant dimensions in all of the present cases because of the error-driven learning mechanism. By shifting the mean of a cluster along a relevant dimension, the model can maximize its ability to discriminate between categories that are separated along this dimension. However, the model, in its current formulation, is unable to capture the distortions toward the center of the category distribution observed on the irrelevant dimensions in the unidimensional conditions. Instead, the model can only predict no distortion along these dimensions because they don't contribute to the task goal. Adding a bias term to reflect this tendency of subjects to distort toward the center of the category distribution would allow the model to predict these effects, and would be psychologically motivated given the findings from the literature on memory for stimulus magnitude discussed above.

General Discussion

The present experiment and simulations show that goals can cause idealization in artificial categories learned in the laboratory. When subjects only were required to discriminate between categories using a single dimension, their reconstructions scores showed that the graded structure was determined by ideals on only this dimension. However, when subjects learned the categories using both dimensions, ideals determined the graded structure on both dimensions. These results are an a priori prediction of error-driven clustering models that explain these effects as arising from simple learning mechanisms.

One contribution of this research is in highlighting the continuity between category learning in the laboratory and more ecologically based studies of concept use. The present research suggests that it is important to consider the role of simple learning mechanisms in producing idealization effects, which are often described as requiring abstract theoretical knowledge. While we do not suggest that error-driven learning models are able to explain all occasions in which goals have an effect on graded structure, we do believe that they provide an important alternate description that may help to predict performance inside and outside of the laboratory.

As with many findings that show differences in category representations arising from differences in people's goals, these results are problematic for approaches to category learning that do not allow task demands to influence category representations. Anderson's rational model (Anderson, 1991), in particular, only considers the statistics of the learning environment in forming representations, and would not predict any of the observed shifts in graded structure, nor any of the differences in reconstructions between conditions. Still, environmental statistics clearly place important constraints on any categorization problem, and researchers need to consider ways of incorporating the effects of both into their models (cf. Love, 2005).

In conclusion, we demonstrate that manipulating the goals of subjects within a task can cause the graded structure of artificial categories learned in the laboratory to be organized around ideals. This is important for category learning modelers and those who focus on laboratory experiments because it helps to provide continuity between cultural and laboratory research. Like other findings regarding goal related influences on category representations, this also has implications for the field of human categorization as a whole, because it offers additional support for the irreducible influence of category use on category representations. Finally, the modeling described above may help to motivate and inform future research in cross-cultural studies of concept use, and thus open a broader dialog between research programs that are often isolated from each other within the field.

Acknowledgments

This work was supported by AFOSR Grant FA9550-04-1-0226 and NSF Grant 0349101 to Bradley C. Love

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.

Barsalou, L.W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure of categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition 11*, 629-654.

Goldstone, R. L. (1996). Isolated and Interrelated Concepts. *Memory & Cognition, 24*, 608-628.

Huttenlocher, J., Hedges, L. V. & Vevea, J.L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*, 1-22.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14*, 195-199.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions and graded category structure among tree experts and novices. *Memory and Cognition, 28*, 41-50.

Sailor, K.M., Miriam, A. (2005). Is memory for stimulus magnitude Bayesian? *Memory & Cognition, 33*, 840-851.